基于电子病历利用支持向量机构建疾病预测模型*

——以重度急性胰腺炎早期预警为例

张　晔[1]　张　晗[1]　尹玢璨[1]　赵玉虹[2]
[1](中国医科大学医学信息学院　沈阳　110122)
[2](中国医科大学附属盛京医院　沈阳　110004)

摘要：【目的】为构建疾病预测模型，以重度急性胰腺炎早期预警为例，提出一种基于支持向量机的疾病预测模型构建方法。【方法】基于支持向量机 LIBSVM3.11，采用优化后的径向基核函数产生的分类器，同时结合统计学单因素及多因素 Logistic 回归分析方法，进行特征变量选取，提出一种简单易行的重度急性胰腺炎早期预警模型。【结果】所构建重度急性胰腺炎预警模型准确率达 70.37%。最终纳入模型变量包括白细胞计数、血清钙离子、血清脂肪酶、收缩压、舒张压及胸腔积液。【局限】样本量有限，主要采用支持向量机构建疾病预测模型，未来可建立系统，突出临床应用价值。【结论】支持向量机可构建疾病预测的最优模型，进一步建立系统，辅助临床决策。

关键词：支持向量机　重度急性胰腺炎　预警　临床决策
分类号：TP393　G35

1　引　言

电子病历(Electronic Medical Record, EMR)即基于计算机的病人记录，是对医疗数据进行电子化保存、管理、传输和重现，主要包括门诊 EMR、住院 EMR、急诊 EMR，其中住院 EMR 由病历首页、入院记录、病程记录、手术(医嘱)记录单、检查报告单等组成[1]。基于 EMR 可获得准确、完整的医疗资料，提示和警示医疗人员，提供临床决策支持。电子病历的核心价值在于临床决策支持，即应用统计分析、数据挖掘等方法，辅助临床决策，对疾病早期预警或特定结局事件发生监测起到重要作用。

随着医疗卫生信息化的发展，电子病历辅助临床决策功能需求增多。建立临床决策支持系统，可减少临床实践过程中误诊或漏诊的出现，同时还可减少医疗资源占用，解决医疗拥挤等问题。

2　相关研究

目前多数临床决策支持应用研究包括疾病诊断、危险因素或复发与否等预测。例如：心力衰竭诊断金标准的制定[2]、阿尔兹海默病进展预测[3]、心肺骤停或死亡事件发生预测[4]以及传染病症状监测系统的创建[5]等。虽然此类研究近年来发展速度较快，然而多数目的在于制定临床标准，或者是已有预测方法的比较评估及新预测方法的提出，并没有完全与临床实际应用接轨。真正的辅助临床决策，不仅仅是建立预测模型或评判预测方法，而是在于如何提高医生工作质量，例如缩短诊疗时间、避免过度医疗、减少医疗差错等。

常用的决策方法有机器学习、统计分析及规则归纳法等，机器学习以支持向量机(Support Vector

通讯作者：赵玉虹，ORCID：0000-0003-4265-6692，E-mail：joan@mail.cmu.edu.cn。

*本文系教育部人文社会科学研究青年基金项目“基于语义述谓网络属性的多文档自动摘要：以生物医学为例”(项目编号:13YJC870030)的研究成果之一。

Machine, SVM)、人工神经网络为主。其中, Kim 等[6]分别基于人工神经网络和 SVM 建立晚期前列腺癌术前预测模型; Kim 等[7]基于 SVM 建立乳腺癌复发预测模型; 吕奕等[8]基于 SVM 提出一有效的肠癌肝转移预测模型。支持向量机是由 Vapnik 等提出的一套学习算法, 是寻找稳健分类模型的一种代表性算法[9]。针对不同领域、不同类型数据(医学数据、金融数据、生物数据等), 有研究者把 SVM 和其他分类方法进行预测模型的性能比较, 结果发现 SVM 算法在分类性能、泛化能力、建模计算量等方面具有明显优势[10]。SVM 的基本思想是定义最优线性超平面, 进而基于 Mercer 核展开定理, 通过非线性映射ϕ, 把样本空间映射到一个高维乃至于无穷维的特征空间, 使在特征空间中可以应用线性学习机的方法解决样本空间中的高度非线性问题[9]。近年来, 随着 SVM 核函数的嵌入, 医学领域中越来越多的临床决策研究者开始应用 SVM 构建预测模型。疾病预测模型建立本质上是一个分类问题。疾病预测分类具有样本数量有限、维数较高等特点, 原理上符合 SVM 的适用条件, 因为 SVM 作为一种专门针对小样本的算法, 既不同于基于大数定律的传统统计方法, 同时也不同于其他机器学习方法如人工神经网络等要求大样本数据, 且会出现网络结构难以确定、过学习、欠学习、学习时间长等问题。本研究基于电子病历首次病程记录和实验室检验数据, 应用支持向量机构建重度急性胰腺炎疾病预测模型, 旨在提出一种简单易行的早期预警模型, 以期进一步建立系统, 辅助临床决策。

3 研究框架与方法

疾病预测主要涉及疾病诊断、进展或复发等预测, 实质上就是"是"与"否"的二分类问题。基于支持向量机构建疾病预测模型具体流程如图 1 所示:

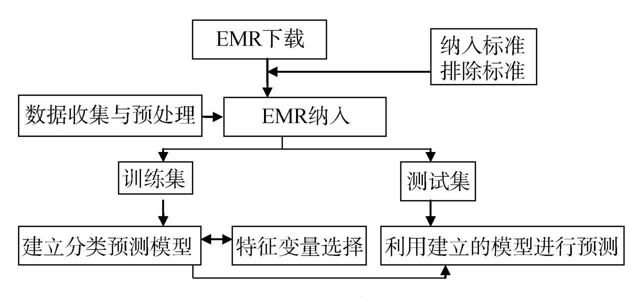


图 1　基于支持向量机构建疾病预测模型流程

(1) 数据下载收集: 确定研究变量及结局变量指标。根据 EMR 首页疾病 ICD 编码, 查询下载相应疾病 EMRs, 下载内容包含首次病程记录、实验室检验报告等。而后根据纳入排除标准纳入符合标准的病历。将纳入病历分为训练集、测试集。

(2) 数据预处理: 如若病历下载内容包括首次病程记录, 则需对文本进行汉语分词处理, 可应用中国科学院计算技术研究所的 NLPIR 软件[11]对各 EMR 首次病程记录的症状特征词初步分词; 经初步分词后, 如若需要, 可基于规则和专业词典针对词义否定识别进行分词调整; 调用 NLPIR 软件程序抽取所调整症状特征词(关键词), 核心代码如下所示:

```
{
int nCount = NLPIR_ImportUserDict("userdic.txt");
NLPIR_FileProcess("Test.txt", "Result.txt",1);
sResult=NLPIR_GetFileKeyWords("test.txt");
printf("Keywords are:\n%s\n",sResult);
}
```

(3) 疾病预测模型建立: 基于 MATLAB2010a 平台, 选用 LIBSVM3.11 工具箱[12]进行支持向量机分类建模。调用函数 xlsread 将数据集 Excel 文件读入转换成所要求的矩阵格式, 如图 2 所示:

```
<label1> <index11>:<value11> ... <index1m>:<value1m>
<label2> <index21>:<value21> ... <index2m>:<value2m>
   .         .         .          .          .
   .         .         .          .          .
   .         .         .          .          .
<label n> <index n1>:<value n1> ... <index nm>:<value nm>
```

图 2　LIBSVM 所需数据格式

其中 label 为样本结局变量标签, index 为研究变量, value 为变量属性值。调用函数 mapminmax 将数据归一化至[–1,1], 以统一变量量纲, 同时简化计算。

SVM 解决非线性问题关键在于引入核函数类型及参数的选择。通过调整参数 t, 调整核函数类型(包括线性、多项式、径向基、sigmoid 函数); 采用网格搜索寻找最优核参数(c, g); 采用最优参数进行分类模型训练, 依据留一法计算预测模型准确率, 以高者为优, 确定最终预测模型及函数表达式。核心代码如下:

```
>> bestcv = 0;
for log2c = -5 : 5,
for log2g = -5 : 5,
cmd = ['-v 5 -c ', num2str(2^log2c), ' -g ', num2str(2^log2g)];
cv = svmtrain(aptr_label, aptr_inst1, cmd);
if (cv >= bestcv),
```

```
bestcv = cv;
bestc = 2^log2c;
bestg = 2^log2g;
end
fprintf('%g %g %g (best c=%g, g=%g, rate=%g)\n', log2c, log2g,
  cv, bestc, bestg, bestcv);
end
end
>> model = svmtrain(aptr_label, aptr_inst1, '-t   -c   -g   ');
>> [predict_label,accuracy]=svmpredict(apte_label,apte_inst,model);
```

(4) 特征变量选取：依据数据分布情况选用不同方法进行单因素统计学分析(独立样本 T 检验或双样本 Kolmogorov-Smirnov 检验或卡方检验)初步筛选变量；而后将 P 值小于 0.2 的变量纳入 Logistic 回归分析，选取最终高预测性能变量，$P<0.05$ 具有统计学意义。

(5) 疾病预测模型再建立：依据所选取特征变量，再次基于 SVM 构建疾病预测模型，比较预测性能前后是否有所提高。

4　研究过程

4.1　分类模型建立

(1) 研究对象

随机抽取辽宁省某医院去识别化的 2013 年 1 月至 2015 年 3 月主诊断为急性胰腺炎(Acute Pancreatitis, AP)的电子病历 323 例，其中非重度 203 例、重度 120 例。纳入标准为：住院病例；出院诊断中主诊断为急性胰腺炎；腹痛发作时间小于 30 天。排除标准为：转院、二次入院病例；慢性胰腺炎、胰腺肿瘤；所研究变量数据不全。

(2) 研究变量

参考 Up To Data 临床顾问数据库[13]中列出的 AP 严重程度危险因素，最终选取的研究变量(n=20)包括：年龄、白细胞计数、红细胞比积、尿素、肌酐、K^+、Na^+、Ca^{2+}、血清淀粉酶、血清脂肪酶、体温、心率、呼吸频率、血压(收缩压/舒张压)、腹痛发作时间，以及是否神清、是否器官衰竭、是否胰腺坏、是否有胸腔积液存在。

(3) 结局变量

根据国际 AP 专题研讨会最新修订的 AP 分级和分类系统，AP 可分为三级：轻度 AP(Mild Acute Pancreatitis, MAP)、中度 AP(Moderately Severe Acute Pancreatitis, MSAP)、重度 AP(Severe Acute Pancreatitis, SAP)[14]。本研究选用出院诊断重度急性胰腺炎与否作为结局变量。

(4) 数据预处理

研究中所纳入变量包括 4 项分类变量，即“是否神清”、“是否器官衰竭”、“是否胰腺坏死”、“是否存在胸腔积液”，以文字描述形式在首次病程记录中出现，需进行汉语分词处理。将纳入样本按 3：1 分为训练集(n=242)和测试集(n=81)，针对训练集，NLPIR 软件初步分词处理各 EMR 首次病程记录的病例特点，包括：现病史、既往史、体格检查辅助检查。需向 NLPIR 分词软件中添加用户词典，包括：ICD-10 编码对应术语，中文 MeSH 主题词(款目词)，中文数据库中相应主题词、关键词，对添加的预测指标词进行关键词标记(Key)；经初步分词后，基于规则(人工制定)和专业词典(AP 诊治标准中标准术语)针对词义否定识别进行分词调整；调用 NLPIR 软件抽取所调整疾病严重程度特征词(描述上述 4 项分类变量)，标记诊断结果(是否 SAP)；针对测试集：初步分词、分词调整、特征词抽取步骤细节除 NLPIR 软件添加用户词典外，其余相同。测试集中向 NLPIR 用户词典中另外添加训练集中所抽取的特征词。

(5) 建立分类预测模型

基于训练集数据建立上述20个变量与结局变量间的分类预测模型，再基于测试集计算所建立模型预测准确率。本研究选用默认参数值，分别选择不同的核函数进行分类建模。预测性能如表1所示：

表 1　不同核函数分类预测性能

核函数	c	g	v	sv	bsv	trA	teA
线性	1	/	242	179	171	69.83%	67.90%
多项式	1	0.05	242	182	178	60.33%	69.14%
径向基(RBF)	1	0.05	242	183	178	59.09%	70.37%
sigmoid	1	0.05	242	184	176	62.81%	62.96%

(注：v：交叉验证折数；sv：支持向量数；bsv：边界支持向量数；trA：训练集准确率，teA：测试集准确率。)

选用径向基核函数，结合网格搜索和交叉验证方法选择最优参数，预测性能如表 2 所示：

表 2　径向基参数优化前后预测性能

方法	c	g	v	sv	bsv	trA	teA
未优化	1	0.05	242	183	178	59.09%	70.37%
网格搜索	16	0.0625	242	175	159	69.01%	67.90%

(注：v：交叉验证折数；sv：支持向量数；bsv：边界支持向量数；trA：训练集准确率，teA：测试集准确率。)

(6) 特征变量选取

对纳入研究变量进行特征变量选取，确定高预测性能变量。采用 SPSS 19.0 进行统计学分析。变量筛选结果如表 3 所示：

表 3　单因素分析结果

变量	缩写	是否 SAP t(p)	是否 SAP z(p)	是否 SAP χ^2 (p)
年龄	Y	0.313(0.755)	/	/
白细胞计数	WBC	–4.784(0.000)*	/	/
红细胞比积	HCT	–0.528(0.598)	/	/
钾	K+	0.114(0.909)	/	/
收缩压	SBP	–1.342(0.181)*	/	/
尿素	Urea	/	1.425(0.035)	/
肌酐	Cr	/	0.518(0.951)	/
钠	Na+	/	1.164(0.133)*	/
钙	Ca2+	/	2.972(0.000)*	/
血清淀粉酶	AMY	/	1.270(0.080)*	/
血清脂肪酶	LPS	/	2.043(0.000)*	/
体温	T	/	1.317(0.062)*	/
心率	P	/	2.043(0.000)*	/
呼吸频次	R	/	0.800(0.544)	/
舒张压	DBP	/	1.879(0.002)*	/
腹痛发作时间	/	/	1.133(0.153)*	/
器官衰竭	OF	/	/	**(0.018)*
精神状态	/	/	/	0.581(0.446)
胰腺坏死	PN	/	/	**(0.372)
胸腔积液	PE	/	/	14.238(0.000)*

(注：若两类样本数据为连续型变量且均服从正态分布，单因素分析选用 T 检验；若两类样本数据为连续型变量但不均服从正态分布，单因素分析选用 Kolmogorov-Smirnov 检验；若两类样本数据为分类变量，单因素分析选用卡方检验；*：单因素分析 $p<0.2$，纳入 Logistic 回归分析；**：采用 Fisher's 精确检验。)

经单因素分析纳入 Logistic 回归分析的变量有白细胞计数(WBC)、收缩压(SBP)、尿素(Urea)、钠(Na^+)、钙(Ca^{2+})、体温(T)、心率(P)、舒张压(DBP)、血清淀粉酶(AMY)、血清脂肪酶(LPS)、腹痛发作时间、器官衰竭(OF)、胰腺坏死(PN)、胸腔积液(PE)。经 Logistic 回归分析，最终纳入特征变量有：白细胞计数、血清钙离子、血清脂肪酶、收缩压、舒张压以及是否伴有胸腔积液。Logistic 回归方程如公式(1)所示：

$$P=1/\{1+\exp[-(4.767+0.126'WBC'-3.142'CA'+0.001'LPS'-0.027'SBP'+0.059'DBP'-2.157'PE')]\} \quad (1)$$

(7) 分类预测模型再建立

选用径向基核函数，结合网格搜索和交叉验证方法选择最优参数。c 最优值为2，g 最优值为1，模型支持向量数为180，训练集和测试集准确率分别为65.29%、70.37%。最终决策函数如下：

$$Predict_y = sign\left(\sum_{i=1}^{n} W_i \exp(-gamma \| x_i - x \|^2) - 0.388\right) \quad (2)$$

其中，$\|x_i-x\|^2$为二范数距离，n 代表支持向量的个数即180；对于每一个 i：w_i=model.sv_coef(i)，即支持向量的系数，x_i=model.SVs(i,:)，即支持向量矩阵。X 是待预测样本，gamma 是参数 g。

4.2　分类模型预测性能评价

采用测试集数据对预测模型的准确率进行客观评估，其中 SAP 类设为正类、非 SAP 类设为负类，使用准确率衡量模型预测性能。计算公式如下：

$$准确率(A)=(真正+真负)/总样本数 \quad (3)$$

本研究测试集 81 个样本中，正类样本数 30，负类样本数 51，其中真正样本数 14，假负样本数 16，假正样本数 8，真负样本数 43。计算得准确率(A)为70.37%。

4.3 对比实验

另选用 Logistic 回归分析、决策树和人工神经网络方法进行对比实验。训练集、测试集样本同 SVM 建立模型所用。采用 WEKA3.7.13 软件包实现上述算法, WEKA 是一基于 Java 语言编写的数据挖掘机器学习软件, 包括完整的数据处理工具、学习算法和评价方法[15]。就分类问题, 可选择不同分类算法(贝叶斯、决策树、多层感知器、支持向量机等)建立不同分类器。本文分别选择 “Classifier” 下 “Logistic” 、 “J48” 、 “Multilayer Perceptron”实现 Logistic 回归分析、决策树和人工神经网络算法, 选择默认参数, 对于训练集同样采用留一法建立模型。最终通过分类模型的评估量化度量模型预测性能。三种方法与 SVM 建模后筛选特征变量后再建模预测准确率比较如表 4 所示:

表 4 不同方法预测准确率比较

方法	v	trA	teA
Logistic	242	95.87%	62.96%
决策树	242	99.59%	62.96%
人工神经网络	242	97.52%	64.20%
SVM	242	65.29%	70.37%

(注: v: 交叉验证折数; trA: 训练集准确率, teA: 测试集准确率。)

5 结果及讨论

由表 1 可知, 不同核函数所对应的分类模型的预测准确率存在一定的差别, 针对测试集数据, 本实验中多项式核函数和径向基核函数的预测准确率较高, 且支持向量的个数较为理想。本实验选用 SVM 构建重度急性胰腺炎早期预警模型, 是由于 SVM 基于核函数实现升维, 可解决非线性分类和回归问题; SVM 的最终决策函数只由少数的支持向量决定, 计算的复杂性取决于支持向量数, 而不是纳入的研究变量数; 相较于其他方法, 建立 SVM 模型所需的人为干预少, 可保证模型的客观性, 因此 SVM 常用来解决医疗数据分类和回归建模问题。应用 SVM 解决非线性问题多选用径向基函数作为核函数。原因是其适合非线性关系; 其模型复杂度较好, 优于多项式; 其数值计算易实现。选用网格搜索和交叉验证的参数寻优方法, 虽然所得结果不一定是理论上的最优, 但也能够满足一定条件下的最优。本实验选用径向基核函数, 并结合网格搜索和交叉验证方法选择最优参数。由表 2 参数优化前后对比可知, 参数优化不仅能够提高模型预测准确率, 同时还可以减少支持向量个数, 从而简化预测函数。

基于 SVM 所建立的 SAP 预测模型, 其预测准确率较为理想, 说明支持向量机方法可用于建立疾病预测模型, 通过核函数选择和参数寻优可以对模型预测性能进行优化。支持向量机所建立模型虽具有较好预测性能, 但在临床实际中模型中变量不一定都与预测结局高度相关甚至相关, 因此模型建立后特征变量选择尤为重要[16]。

根据数据分布不同针对各分布数据选择不同单因素分析方法, 目的在于初步筛选变量, 剔除非高度相关变量或者某些同效变量, 同时减少进入 Logistic 回归分析所需样本数。所选变量具体从 AP 炎性反应、特征酶变化及伴发症影响 AP 严重程度进展。基于所选变量再建预测模型, 预测模型准确率较前提高。说明基于特征选取简化数据维数, 不仅能够摆脱与预测任务不相关的数据、显著减少所需的训练集样本数量, 同时还能够提高模型预测性能。

在对比实验中, 如表 4 所示, Logistic 回归分析、决策树和人工神经网络三种方法, 训练集准确率较 SVM 高, 但测试集准确率均较 SVM 低。原因在于此三种方法在构建模型时, 以经验风险最小为原则, 可能存在过拟合的现象。疾病预测模型的建立最终问题是寻找稳健分类模型, SVM 作为一种以结构风险最小化原理为基础的算法, 权衡训练样本的平均预测误差与模型的复杂度, 以经验风险和置信区间的和最小为目标, 所建立的分类模型具有较好的鲁棒性。因此测试集准确率较为理想。

6 结 语

随着医院信息化的建设, 电子病历其核心价值即临床决策支持将成为未来发展的方向。本研究基于 SVM, 以重度急性胰腺炎为例, 基于电子病历构建疾病预测模型。本研究特点包括: 以重度急性胰腺炎为例, 尝试采用文字型及数值型医疗数据建立疾病早期预警模型, 以期进一步建立决策系统; 选用支持向量机建立预测模型后结合统计学分析方法筛选特征变量, 而后再基于特征变量建立最终预测模型, 提高预

测准确率的同时简化模型。研究也存在不足，表现在临床应用方面。在今后的研究中，将进一步根据所建立模型创建决策系统，突出其临床应用价值。此外，以临床需求为出发点，增加样本数，应用决策方法构建疾病预测模型，并且如何服务于临床，也是从事临床决策支持研究者应共同努力的方向。

参考文献：

[1] 雷健波. 电子病历的核心价值与临床决策支持[J]. 中国数字医学, 2008, 3(3): 26-30. (Lei Jianbo. Clinical Decision Support and the Core Value of Electronic Medical Record [J]. China Digital Medicine, 2008, 3(3): 26-30.)

[2] Byrd R J, Steinhubl S R, Sun J, et al. Automatic Identification of Heart Failure Diagnostic Criteria, Using Text Analysis of Clinical Notes from Electronic Health Records [J]. International Journal of Medical Informatics, 2014, 83(12): 983-992.

[3] Ye J, Farnum M, Yang E, et al. Sparse Learning and Stability Selection for Predicting MCI to AD Conversion Using Baseline ADNI Data [J]. BMC Neurology, 2012. DOI: 10.1186/1471-2377-12-46.

[4] Alvarez C A, Clark C A, Zhang S, et al. Predicting out of Intensive Care Unit Cardiopulmonary Arrest or Death Using Electronic Medical Record Data [J]. BMC Medical Informatics and Decision Making, 2013. DOI: 10.1186/1472- 6947-13-128.

[5] Matheny M E, Fitzhenry F, Speroff T, et al. Detection of Infectious Symptoms from VA Emergency Department and Primary Care Clinical Documentation [J]. International Journal of Medical Informatics, 2012, 81(3): 143-156.

[6] Kim S Y, Moon S K, Jung D C, et al. Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Artificial Neural Network [J]. Korean Journal of Radiology, 2011, 12(5): 588-594.

[7] Kim W, Kim K S, Lee J E, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine [J]. Journal of Breast Cancer, 2012, 15(2): 230-238.

[8] 吕奕, 王清. 一种基于概率校正和集成学习的肠癌肝转移预测模型[J]. 计算机应用与软件, 2011, 28 (9): 48-51. (Lv Yi, Wang Qing. A Probability Calibration and Ensemble Learning Based Colorectal Cancer Liver Metastasis Prediction Model [J]. Computer Applications and Software, 2011, 28 (9): 48-51.)

[9] 王星, 等. 大数据分析: 方法与应用[M]. 北京: 清华大学出版社, 2013: 68-90. (Wang Xing, et al. Big Data Analysis: Methods and Applications[M]. Beijing: Tsinghua University Press, 2013: 68-90.)

[10] 陈永义, 熊秋芬. 支持向量机方法应用教程[M]. 北京: 气象出版社, 2011: 6-10. (Chen Yongyi, Xiong Qiufen. Application of Support Vector Machines Tutorial [M]. Beijing: China Meteorological Press, 2011: 6-10.)

[11] ICTCLAS 2014 [EB/OL]. [2015-03-25]. http://ictclas.nlpir.org/.

[12] LIBSVM—A Library for Support Vector Machines [EB/OL]. [2015-03-25]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[13] Up To Data [EB/OL]. [2015-03-25]. http://www.uptodate.com/contents/search.

[14] Banks P A, Bollen T L, Dervenis C, et al. Classification of Acute Pancreatitis--2012: Revision of the Atlanta Classification and Definitions by International Consensus [J]. Gut, 2013, 62(1): 102-111.

[15] 袁梅宇. 数据挖掘与机器学习——WEKA 应用技术与实践[M]. 北京: 清华大学出版社, 2014: 2. (Yuan Meiyu. Data Mining and Machine Learning——WEKA Application Technology and Practice [M]. Beijing: Tsinghua University Press, 2014: 2.)

[16] 刘勘, 朱怀萍, 刘秀芹. 基于支持向量机的网络伪舆情识别研究[J]. 现代图书情报技术, 2013(11): 75-80. (Liu Kan, Zhu Huaiping, Liu Xiuqin. Detection of Internet Deceptive Opinion Based on SVM [J]. New Technology of Library and Information Service, 2013(11): 75-80.)

作者贡献声明：

张晔：设计研究方案，负责实验，采集、清洗和分析数据，论文起草及最终版本修订；

张晗：提出研究思路，修改论文；

尹玢璨：采集、清洗和分析数据。

赵玉虹：提出研究思路，修改论文。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，可通过电子邮件向作者索取，E-mail: 1332457636@163.com。

[1] 张晔，赵玉虹. ap.xls. 急性胰腺炎电子病历纳入样本变量数据.

[2] 张晔，赵玉虹. apdic.txt. 急性胰腺炎专业词典.

收稿日期: 2015-09-21

收修改稿日期: 2015-12-07

Building Disease Prediction Model Using Support Vector Machine ——Case Study of Severe Acute Pancreatitis

Zhang Ye[1] Zhang Han[1] Yin Bincan[1] Zhao Yuhong[2]
[1] (Department of Medical Informatics, China Medical University, Shenyang 110122, China)
[2](Shengjing Hospital of China Medical University, Shenyang 110004, China)

Abstract: [**Objective**] This study developed a disease prediction model based on the support vector machine, using electronic medical records of the severe acute pancreatitis patients. [**Methods**] We first adjusted the kernel type and parameter values of the support vector machine method to get an optimized prediction model. Then, we combined it with univariable and multivariable logistic regression analysis methods to select features' variable. Finally, we proposed a simplified early warning model for the severe acute pancreatitis. [**Results**] The new model's prediction accuracy rate is 70.37%. Variables used by this model include: white blood cell count, serum calcium, serum lipase, systolic blood pressure, diastolic blood pressure and pleural effusion. [**Limitations**] Because of the small sample size, we only used this support vector machine method to develop the new disease prediction model. In the future, we will try to establish a larger examination system for the clinical trial. [**Conclusions**] Support vector machine can help us develop an optimal disease prediction model. A new system based on this model could support our clinical decision makings.

Keywords: Support Vector Machine Severe acute pancreatitis Early warning Clinical decision

---